Exploring the forecasting approach for road accidents: Analytical measures with hybrid machine learning

Mamoudou Sangare, Sharut Gupta, Samia Bouzefrane, Soumya Banerjee, Paul Mühlethaler

Exploring the Forecasting Approach for Road Accidents: Analytical measures with Hybrid Machine Learning

Mamoudou Sangare[a,b] , Sharut Gupta[a] , Samia Bouzefrane[b] , Soumya Banerjee[b] , Paul Muhlethaler[a]
[a]Inria, Paris, France
[b]Cnam, Paris, France

Abstract

Urban traffic forecasting models generally follow either a Gaussian Mixture Model (GMM) or a Support Vector Classifier (SVC) to estimate the features of potential road accidents. Although SVC can provide good performances with less data than GMM, it incurs a higher computational cost. This paper proposes a novel framework that combines the descriptive strength of the Gaussian Mixture Model with the high-performance classification capabilities of the Support Vector Classifier. A new approach is presented that uses the mean vectors obtained from the GMM model as input to the SVC. Experimental results show that the approach compares very favorably with baseline statistical methods.

Keywords: Gaussian mixture model, Support vector machine, Support vector classifier, Hybrid model, Traffic accident forecasting, Accident severity

Conflicts of interest :

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honorary; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names: Mamoudou Sangaré , Sharut Gupta , Samia Bouzefrane , Soumya Banerjee, Paul Muhlethaler

1. INTRODUCTION

It is a regrettable fact that the number of road traffic accidents continues to rise, largely due to rapid urban growth and the ever-increasing density of vehicles in cities and surrounding areas. According to statistics from World Health Organization, each year approximately 1.25 million people lose their lives in road traffic accidents worldwide, which means that one person is killed every 25 seconds. The statistics predict that road accidents will grow by 65% and will become the fifth greatest cause of fatalities by 2030. However, with emerging sensor devices and the IoT, it has become feasible to configure future vehicles with safety sensors to prevent many of these accidents.

Therefore, recent research has been orchestrated to investigate the state-of the-art of vehicle safety analysis in Vehicular ad-hoc networks (VANETs) while including the issue of road accidents. It has become clear that certain driving safety [1] and road safety issues involve various dimensions and parameters. For example, in driving safety analysis, the style and behavior of drivers must be studied to investigate unorthodox driving style and neighboring conditions are sensed and analyzed in order to provide intelligent support for the drivers. However, the major constraints of such analyses are the absence of substantial data sets and a precise means of identifying the neighboring conditions without incorporating additional sensors.

Another aspect of road safety analysis involves the effect on road safety from external parameters, including the road surface, geometry, traffic flow, weather conditions and both drivers' and pedestrians' behavior. Despite considerable research efforts, it has not been possible to provide the most deterministic and computationally intelligent model [2] to predict the exact context of road accidents due to unbalanced data instances at various levels. Accident prediction is one of the most crucial aspects of road safety where precautionary measures are taken to avoid an accident before it occurs.

Therefore, it is important to investigate the accident-prone areas of cities and the effect of external factors in order to be able to forecast the safety level of roads with appropriate granularity. There are huge variations in road traffic accidents in terms of the extremity of the accident and the damage to people and their property, which is also referred to as accident severity [3]. It is necessary to investigate the relationship between traffic accident severity and related risk factors such as the traffic volume, the driver's age, maximum possible speed, geometrical factors like the type of vehicle, distance from the nearest intersection, details of intersections etc. and environmental features such as weather conditions, lighting conditions, type of road, etc. Previous studies in this area are broadly classified into two categories: statistical modeling and machine learning modeling.

Initially, accident severity analysis for traffic accident forecasting was primarily carried out using statistical techniques [49, 50, 51, 52, 53]. Statistical models are favored over machine learning models due to their solid theoretical base and strong mathematical formulation [47]. Gaussian Mixture Models (GMMs) are the most common, and they have been successfully used in a wide variety of fields, such as voice recognition systems, video image processing, and pattern classification. In studies related to traffic flow and accidents, GMMs have been repeatedly used to model raw time-to-collision (TTC) samples for traffic safety prediction [6, 7], severity detection of traffic accidents along with Hidden Markov models [10] and in traffic flow forecasting [11]. However, these models assume some inherent properties about the data patterns like the assumption of risk factors influencing accident severity linearly, which might not always be the case [46, 22] and hence they inevitably induce inaccurate results.

Machine learning models on the other hand [48], are highly adaptable with no or very few assumptions about the input features and offer higher flexibility to outliers, as well as inaccurate and missing data. Popular machine-learning models applied to traffic accident related studies include Decision Trees [34, 35], Support Vector Machines (SVMs) and Support Vector Classifiers (SVCs) [43, 44, 45], K-means clustering [41, 42], Artificial Neural Networks (ANNs) [36, 37, 38] etc. Of these models, Support Vector Machines have been increasingly used in traffic related studies to address traffic flow prediction [23, 24, 25, 26], crash frequency analysis [27, 28, 29, 30], and to

analyze accident severity in a crash [44, 31, 32]. The major drawback of machine-learning models is their performance as a 'black-box' which leads to unclear inference of the function that correlates the input variables with the target class [33].

The purpose of this study is to combine the Gaussian Mixture Model from statistical modeling and the Support Vector Classifier from machine-learning modeling to overcome the disadvantages of one model by the advantages of the other and hence improve the overall accuracy. Favored by its innate discriminative power, even in the case of non-linearly separable classes using kernels, the SVM presents an attractive way of enhancing the baseline generative model (GMM) [9]. Hence the GMM serves as a parametric basis for the Support Vector Classifier. Since SVCs perform poorly on unbalanced data and cannot select relevant attributes with respect to the target variable, data pre-processing using re-sampling techniques and feature importance ranking methods are applied.

The rest of the paper is organized as follows:
Section 2 introduces specific and recent related research on road safety prediction incorporating various levels of machine learning and intelligent algorithms. The third section gives a brief explanation about the dataset used. Section 4 contains a comprehensive description of the model construction mechanism along with model specifications, followed by the details of the various models used. Section 4.1.1 presents the data pre-processing and balancing techniques used in the proposed model. In Section 5, the results are given with a comparison with the baseline model and data analysis is performed to strengthen the claims of the paper. Finally, the last section concludes the paper with a brief summary of the research limitations and the overall research effort.

2. Review of the literature

In this section we review road safety prediction methods that incorporate various levels of machine learning and intelligent algorithms. Then, we analyze the advantages and drawbacks of various methods, and formulate the missing value problem, which is a substantial challenge in this area of research.

For traffic flow forecasting, various machine learning methods have been employed. The most prominent among these are: Autoregressive Integrated Moving Average (ARIMA) which belongs to time series categories [65]; probabilistic graphical models, such as Bayesian Network [66], Markov Chains [67], and Markov Random Fields (MRFs) [70]; and nonparametric approaches, such as Artificial Neural Networks (ANNs) [67], Support Vector Regression (SVR) [69] , and Locally Weighted Learning (LWL) [70] . However, as seen in the literature, there are multiple reasons for fluctuation in the traffic flow. In addition to that, the patterns in the data are multimodal. These multimodal properties make it difficult to learn. Moreover, for these shallow network approaches to be able to model complex mapping, they require a high dimensional space. A high dimensional space request leads to the requirement of a huge amount of annotated data. Therefore, in the high-dimensional space the overfitting problem becomes acute. In order to overcome this issue, we use a multilayer nonlinear structure since deep learning approaches have a strong ability to express multi-model patterns in data using a reduced number of dimensions. An ANN (Artificial Neural Network) [55] is a type of network in machine learning that has been widely used for road incident prediction in different environments (freeway, highway, urban and non-urban roads, etc.) in order to minimize injury and loss of life on the roads.

An ANN aims to reproduce and simulate human behavior and cognitive functions. It uses a network of nodes, often called neurons, that contain configurable weights and these weights can be trained to produce a desired output [56]. Many kinds of pattern recognition problems can be solved by configuring the layers and the weights of the network. Today, machine-learning techniques have found different applications in a number of fields. These include road safety where they have been used for collision detection. As an illustration, in the study presented in [57] by Chang, an ANN is implemented to predict collisions on a National Freeway in Taiwan. Road features were used as input in their model and it is claimed that the model accepted those features and provided the

number of collisions as output. However, the ANN model has many local minimums, which makes it difficult to find the global optimum solution. This highlights the fact that, even though ANN models are easy to understand, the solution covered by their weights space is non-convex. This is one of its drawbacks. Another shortcoming of an ANN is that it is supervised-based learning and therefore the model requires training data, which limits its applicability in real-world situations. In order to overcome these issues, back propagation (BP) such as Bayesian regularization has been proposed. Although it has been found that Bayesian regularization has led to great improvements, it still requires training data, again making its applicability in the real world limited.

Bayesian Networks (BNs) have also become very popular in traffic prediction as they allow multiple inputs of data to be taken into account. It is known to have applications that can take many forms. It has been pointed out that the inputs of BNs sometimes show less relativity than is the case for neural networks [60][61]. This specific characteristic of BNs offers more possibilities for combining different prediction factors.

Research has revealed that, for traffic prediction, there is no single method that is the best for every situation. Thus, in traffic forecasting, researchers are constantly trying to combine different models. It has been observed that almost all the research undertaken in using hybrid models (HMs) for traffic prediction yield greater prediction accuracy than when a single model is used [62][63]. Jiaming Xie and Yi-King Choi conducted research on designing and implementing a hybrid model that can forecast the traffic flow in the city of Hong Kong by using historical and real-time data. The question that arises here is how one can balance the importance of historical data and real data. This is because it obvious that the traffic situation changes over time and that continuous changes make the traffic status dynamic [64].

As no single model can be suitable for prediction in all situations, the main objective of this research is to build a prediction model that combines two approaches (the Gaussian mixture model and support vector classifiers) in order to predict traffic accidents. The improvement in terms of accuracy is very notable compared to other models.

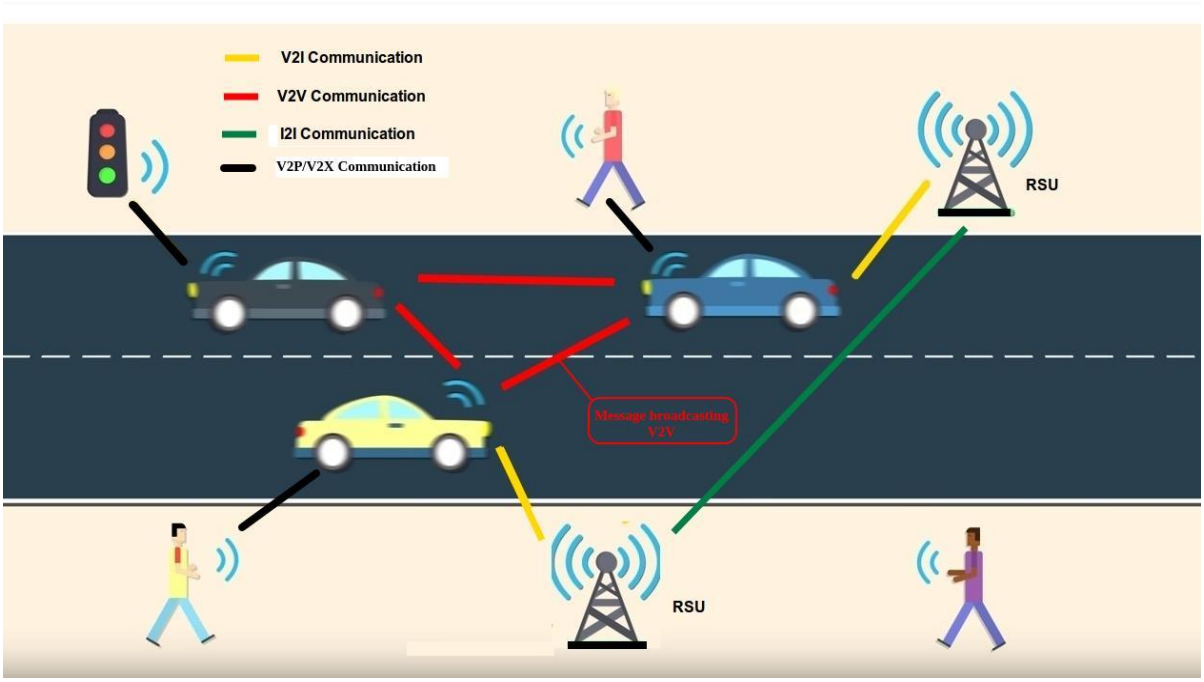

Figure 1: Vehicular ad-hoc networks architecture and message broadcasting scenario

2.1. Background of Vehicular Ad-hoc NETworks (VANETs)

The development of Intelligent Transportation Systems (ITSs) and other applications to improve driving comfort have been motivated by the continuing increase in road traffic accidents.

In order to make these applications feasible, a communication network, the so-called vehicular ad-hoc network (VANET) was developed. In such a network, vehicles are equipped with wireless devices that allow them to communicate. ITS services are provided to the end users by VANETs that transfer data and safety messages. By using wireless standards such as Dedicated Short-Range Communication (DSRC) and Wireless Access in Vehicular Environments (WAVE), VANETs can provide wireless communication between moving vehicles. Essentially, there are three units involved in vehicle communication: Application Units (AUs), On Board Units (OBUs) and Roadside Units (RSUs). By employing wireless standards like IEEE 802.11p [64], communication between OBUs and AUs with RSUs can be possible, as shown in Figure 1.

- Application Units (AUs) - These are smart devices that provide safety applications. They use OBUs to communicate with RSUs. As chips are getting smaller and smaller, these units may be isolated or they may be integrated with OBUs as a single unit. The connection mode of AUs to the OBUs can be done through a wired or wireless connection.

- On Board Units (OBUs) - These are generally installed onboard the vehicles, and their main task is to provide communication between other OBUs and RSUs. In terms of composition, they are made up of devices such as Resource Command Processors (RCP). These resources include a user interface and read/write memory. They also have a specialized interface to connect to other OBUs and a network device for short-range wireless communication based on IEEE 802.11p radio technology. In addition, these units are used for IP mobility management, congestion control, wireless data access, reliable message transfer, data security and geographical routing.

- Roadside Units (RSUs) - These are considered to be fixed nodes that act as a router to provide services to the moving vehicles. RSUs are set up as rigid units at the side of the road in such a way as to maintain coverage and connectivity to all the vehicles. They are the source of radio wave propagation between RSUs and OBUs. The main function of RSUs is to increase the communication range of the Ad Hoc network. They do so by sharing information with OBUs and by sending the information to other RSUs. RSUs work as an information source and provide Internet connectivity to the OBUs. RSUs can be connected to the Internet via a gateway.

VANETs provide the radio interface required by vehicles (wireless transceivers based on IEEE 802.11p, which operate on the dedicated short-range communication (DSRC) band) to communicate with each other using vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2I) and offer a wide variety of applications for ITSs.

As the aim of this work is to derive analytics for the prediction of road accidents, it is important to include a data acquisition mechanism (see Figure 2) and inter-process communication between vehicles on the road. For a standard use case, we consider a segment of a densely populated city road where this type of acquisition model can be placed. To formulate such a model, the physical components of the VANETs can be one of the parameters.

However, the inter-process communication protocol of 5G and beyond can establish a more reliable process exchange mechanism.

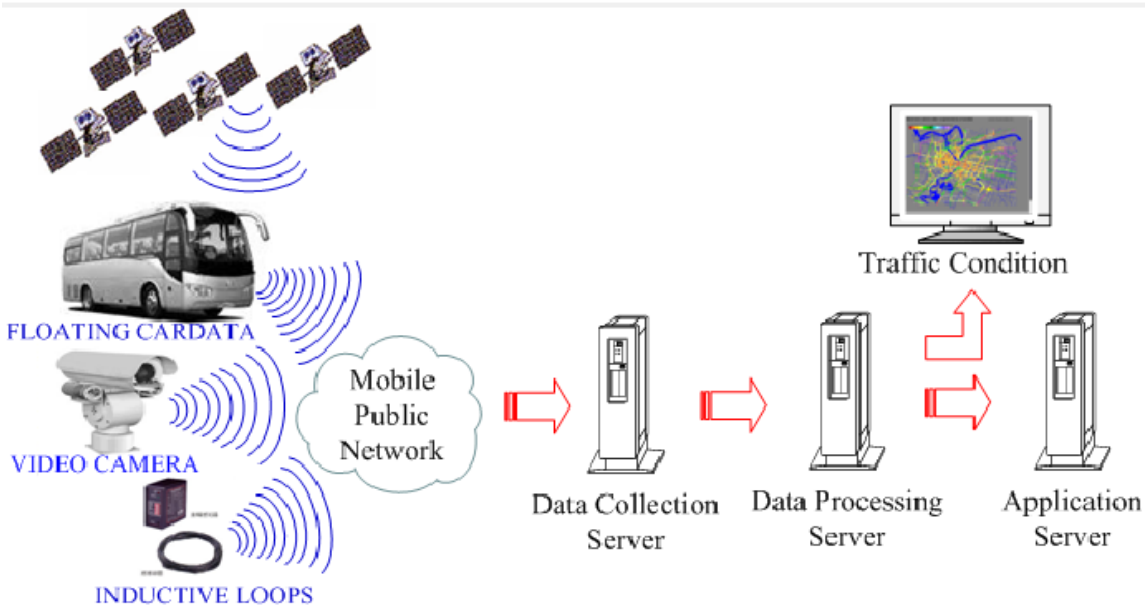

Figure 2: Traffic-data- acquisition -using-different-detectors [72]

In conventional usage, a Message Application Programming Interface (MAPI) is suitable for use cases to collect and to propagate the data for a particular segment of road. In this paper, certain realistic scenarios are considered to formulate the data acquisition and the message passing mechanism.

Four layers are involved in the overall system flow for message passing and broadcasting:

- Application layer (message wrapping mechanism)
- Transport layer (with handshaking between the sender and the receiver)
- Network layer (message distribution mechanism)
- Physical layer (connection and devices)

Thanks to these layers, the system can collect on-road data to pass them towards a nearest cloud center or adjacent vehicle, so that the neighboring vehicle can receive an alert containing several features.

The format of the message is:

1. transaction ID,
2. previous transactions ID, if any
3. sender ID
4. reputation ID (this number will present the reputation or likelihood of a given vehicle to have an accident)
5. receiver ID
6. message content (message text, location, direction of sender vehicle)
7. message type (this information will deliver the warning or the prediction of the message after using a hybrid analytical algorithm such as intersection movement, tendency for forward collision, deviation from the lane, extreme conditions of the road surface and other relevant features).

As well as all these message fields, each message will be followed by the action requested. For example, the action requested indicates the message from the sender vehicle to all receivers for a particular alert if the adjacent vehicle could change lane abruptly. Similarly, if a vehicle accelerates more than normal, despite the traffic congestion on the road, then there will be an alert message for other neighboring vehicles.

All these features of message access and distribution can be implemented in a traffic environment simulator (SIMO) which can be deployed by another event-based simulator like

OMNET++. However, since the objective of our work is to analyze the action requested by the given message distribution format, we did not use these kinds of simulation tools.

2.2. Mathematical symbols at glance

Symbols used in the Gaussian mixture model	
Symbols	**Semantics**
p(x/λ)	Conditional probability of x given
$N_i(x)$	Gaussian probability density function
w_i	Mixing weights
K	Number of Gaussian components
x	Observed data
μ_i	Mean vector of Gaussian components
Σ_i	Co-variance matrix of the Gaussian components
λ	Set of tuples of model parameters w_i, μ_i, Σ_i
M	Number of training examples
L(x/λ)	Log likelihood
Symbols used in the support vector classifier	
Symbols	**Semantics**
(x_i, y_i)	Data examples
f(x)	Classification function
w	Normal direction cosines to the line
$\|\|a\|\|$	Vector norm
ξ_i	Slack variable
C	Penalty index for outliers
α_i	Linear combination weights
K	Kernel
$\langle x, y \rangle$	Inner product between x et y
d	Degree of polynomial kernel
σ	Variance in the RBF kernel

Table I. Mathematical symbols

3. Dataset description

The dataset was obtained from data.govt.uk [18] which is a United Kingdom Government project providing open source data published by central government, local authorities and public bodies. The road traffic accident database for the year of 2017 was used in this research. These dataset files provide detailed road safety data about the environmental, physical, geometrical, geographical and personal information related to accidents as shown in Table II. The data points correspond only to those accidents where information was reported to the police or the authorities. The dataset was compiled from the information recorded in the STATS19 accident reporting form. The entire dataset is mainly composed of three main categories: accident, vehicle and casualty data. The accident variables have 31 features including the weather conditions, lighting, time, day of the week, number of vehicles involved, etc. The Vehicle-Driver database consists of 22 features such as

the age of the vehicle, sex and age of the driver, etc. and the casualty dataset contains 15 feature variables such as the type of victim, sex of the casualty, age of the casualty, etc.

Accident variables	Vehicle and driver variables	Casualty variables
Index of the crash		
Police Force		
Accident Severity		
Vehicles involved		
number of victims involved	Index of the crash	
Date	Vehicle Id code	
Day of the Week	Kind of vehicle	
Time	Towing and Articulation	Index of the crash
Longitude	Vehicle Maneuver	Vehicle Id code
Latitude	Position of vehicle	Casualty Id code
District of Local Authority	Location of intersection	Type of victim
Local Highway Authority	Skidding and Overturning	Gender of victim
1st Road Class	Hit Object in Carriageway	Age of victim
1st Road Number	Vehicle Leaving Carriageway	Age Band of victim
Kind of roadway	Hit Object off Carriageway	Intensity of the fatality
maximum velocity possible	1st Point of Impact	Position of the pedestrian
Details of intersection	Was Vehicle Left Hand Drive	Motion of the pedestrian
Traffic control at intersection	Journey Purpose of Driver	Position of victim in car
2nd Road Class	Gender of the Driver	Position of victim in Bus or coach
2nd Road Number	Age of the Driver	Type of fatality
Pedestrian Crossing-Human Control	Age band of the Driver	Casualty IMD Decile
Pedestrian Crossing-Physical Facilities	Motor power	Location of victim’s home
Lighting	Vehicle fuel type	
Road conditions	Age of the vehicle	
road surface conditions	Rider IMD Decile	
Special characteristics of accident location	Rider Home Area Type	
Carriageway Hazards		
Type of area		
Police attention		

Table II. Dataset variables

The output class or accident severity is divided into 3 categories, namely “no injury in the accident” encoded as 1, “non-incapacitating injury in the accident” encoded as 2 and” incapacitating injury in the accident” encoded as 3. The detailed encoding of every variable is listed in Table III.

Variable name	Variable categories	Code	Frequency	Percentage	Class 1	Class 2	Class 3
Day of the week	Sunday	1	21836	12.20%	20.4%	11.4%	14.7%
	Monday	2	24653	13.77%	12.4%	14.1%	12.2%
	Tuesday	3	25911	14.48%	10.4%	14.7%	13.6%
	Wednesday	4	25837	14.44%	14.4%	14.5%	14.1%
	Thursday	5	27014	15.09 %	11.0 %	15.4 %	14.2%
	Friday	6	29227	16.33%	13.1%	16.5%	15.8%
	Saturday	7	24440	13.66%	18.2%	13.1%	15.3%
Kind of roadway	Traffic circle	1	10328	5.77%	1.05%	3.67%	6.35%
	Single direction traffic	2	2740	1.53%	0.68%	1.28%	1.60%
	Divided highway	3	33739	18.85%	22.9%	16.62%	19.27%
	Undivided highway	6	128608	71.88%	74.8%	77.12%	70.61%
	Slip road	7	1904	1.06%	0.37 %	0.82%	1.13%
	Unknown	9	1599	0.89%	0.06%	0.46%	1.01%

	Single direction traffic/Slip road	12	0	0%	0.00%	0.00%	0.00%
	Erroneous data	-1	0	0%	0.00%	0.00%	0.00%
Human pedestrian crossing	None within 50 metres	0	175819	98.26%	0.12%	0.42%	0.60%
	School crossing patrol	1	497	0.27%	0.06%	0.31%	1.03%
	By another official	2	1013	0.56%	99.5%	99.0%	99.0%
	Erroneous data	-1	1589	0.88%	0.27%	0.26%	0.28%
Details of intersection	No intersection within 20m	0	78468	43.85%	72.01%	49.80%	41.86%
	Traffic circle	1	13524	7.55%	1.33%	4.79%	8.32%
	Mini-Traffic circle	2	1770	0.98%	0.12%	0.68%	1.07%
	Staggered intersection	3	50618	28.29%	15.54%	27.82%	28.67%
	Slip road	5	3502	1.95%	2.32 %	1.52%	2.04%
	Crossroads	6	18403	10.28%	5.05%	8.67%	10.77%
	More than 4 arms (not Traffic circle)	7	1663	0.92%	0.34%	0.60%	1.01%
	Private road	8	5102	2.85%	1.73%	3.28%	2.77%
	Other intersections	9	5567	3.11%	1.52%	2.73%	3.23%
	Erroneous data	-1	301	0.16 %	?%	0.04%	0.20%
Lighting	Daylight	1	126049	70.45%	60.28%	67.95%	71.24%
	Dark with lights	4	36489	20.39%	17.18%	20.02%	20.55%
	Dark with dimmed lights	5	1123	0.62%	0.86%	0.59%	0.63%
	No illumination at all	6	10314	5.76%	20.81%	8.94%	4.70%
	Unknown illumination	7	4941	2.76%	0.83 %	2.48 %	2.86 %
	Erroneous data	-1	2	0.001 %	?%	?%	0.001%
Weather conditions	Breeze	1	144368	80.68%	84.02%	81.45%	80.44%
	Rain with breeze	2	20948	11.70%	8.99%	11.93%	11.71%
	Snow with breeze	3	915	0.51%	0.49%	0.29%	0.56%
	Fine with gale	4	2032	1.13%	1.55%	1.31%	1.08%
	Rain with gale	5	1705	0.95%	1.27 %	1.16%	0.89%
	Snow with gale	6	163	0.09%	?%	0.04%	0.10%
	Fog or mist	7	940	0.52%	0.77%	0.64%	0.49%
	Others	8	3410	1.90%	1.42%	1.55%	1.99%
	Unknown	9	4435	2.47%	1.45%	1.59%	2.70%
	Erroneous data	-1	2	0.001 %	?%	?%	0.001%
Road surface	Dry	1	124151	69.39%%	68.38%	69.25%	69.44%
	Wet	2	49632	27.74%	28.88%	28.55%	27.52%
	Snow	3	707	0.39%	0.31%	0.23%	0.43%
	Icy or snow	4	3172	1.77%	2.23%	1.61%	1.79%
	Flood over 3cm. deep	5	172	0.09%	0.18 %	0.17%	0.07%
	Oily	6	0	0%	0.00 %	0.00 %	0.00 %
	Silt or mud	7	0	0%	0.00%	0.00%	0.00%
	Erroneous data	-1	1084	0.60%	?%	0.16%	0.72%
Sex of the driver	Male	1	114282	63.87%	79.42%	69.89%	62.15%
	Female	2	51848	28.97%	18.33%	24.15%	30.31%
	Not known	3	12784	7.14%	2.23%	5.95%	7.52%
	Erroneous data	-1	4	0.002 %	?%	0.003%	0.002%
Type of victim	Driver	1	121870	68.11 %	63.382%	66.20%	68.65%
	Passenger	2	43204	24.14%	25.00%	22.74%	24.441%
	Pedestrian	3	13844	7.73%	11.60%	11.05%	6.89%
Police officer attend?	Yes	1	143847	80.39 %	95.90%	89.27%	78.02%
	No	2	35055	19.59 %	4.09%	10.71%	21.96%
	No - accident was reported using a self completion form	3	16	0.009 %	?%	0.006%	0.10%
Age band	0-5	1	35	0.02%	?%	0.03%	0.01%
	6 - 10	2	270	0.15%	?%	0.14%	0.15%
	11 - 15	3	1234	0.69%	0.31%	0.71%	0.69%
	16 - 20	4	12705	7.10%	6.48%	7.72%	6.97%
	21 - 25	5	19365	10.82%	9.96 %	11.36%	10.71%
	26 - 35	6	38417	21.47%	19.85%	19.77%	21.89%
	36 - 45	7	29715	16.6%	16.47%	15.76%	16.80%
	46 - 55	8	28832	16.11%	16.81%	16.83%	15.93%
	56 - 65	9	17102	9.55%	12.90%	10.70%	9.22%
	66 - 75	10	8770	4.90%	7.41%	5.95%	4.60%
	Over 75	11	5406	3.02%	6.05%	4.10%	2.70%
	Erroneous data	-1	17067	9.53%	3.73%	6.89%	10.27%
Skidding or overturning	None	0	156030	87.20 %	76.54%	83.00%	88.40%
	Skidding	1	13824	7.72%	14.17%	10.73%	6.89%
	Skidding and overturning	2	3345	1.87%	4.46%	2.86%	1.58%
	Jackknifing	3	52	0.02%	0.12%	0.05%	0.02%
	Jackknifing and overturning	4	56	0.03%	0.21 %	0.02%	0.03%
	Overturning	5	3566	1.99%	4.46%	2.98%	1.71%
	Erroneous data	-1	2045	1.14 %	?%	0.32%	1.35%

Table III. Variable description

4. Research Methodology

4.1. Algorithm description

This section presents the models used in this study for traffic accident forecasting. As mentioned above, the accident data including vehicle, casualty and drivers' features are collected from data.govt.uk. These higher dimensional features are then preprocessed to remove any kind of erroneous entries and balance the dataset. Moreover, if the dataset has a highly unequal distribution of the number of data points corresponding to each class, the SVC model tends to predict every data sample as the majority class. In order to achieve an unbiased performance, it is necessary to balance the dataset with respect to the output class.

Since the dataset is high dimensional, dimensionality reduction techniques must also be used. Furthermore, like Bayesian network (BN) models [4], SVMs lack the ability to automatically select the relevant features. Feature or attribute selection helps to target both above-mentioned disadvantages. Variable importance ranking methods are deployed and the data are further cleaned. This processed dataset is then used as input to the Gaussian Mixture Model [19], [20] and [21] which estimates the parameters of the various Gaussians mixture using expectation maximization. Out of all the parameters i.e. the mean, variance and the mixing probability, the vector of means is adapted and used as input to the SVC model [8]. The SVC treats the accident severity modeling as a classification problem i.e. the accident data is classified into various categories based on the severity classes. This trained hybrid model is then evaluated with respect to the performance metrics and sensitivity analysis is performed. The model is also compared to the baseline GMM model and the results are reported. A brief description of the hybrid algorithm is shown in Figure 3.

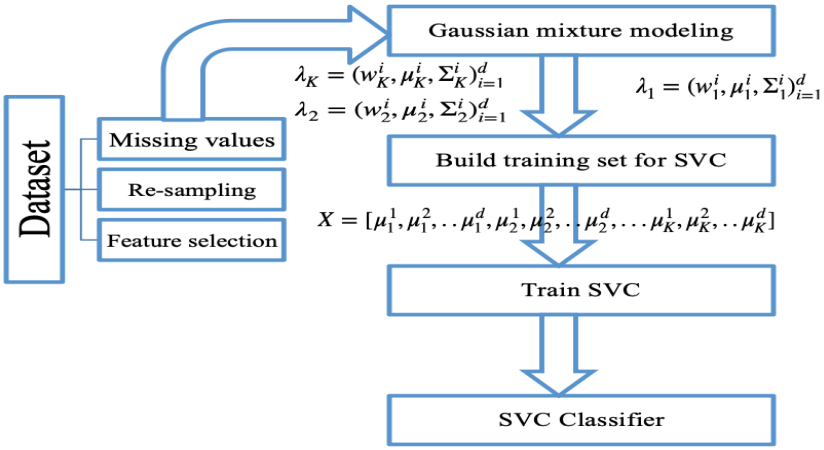

Figure 3. Algorithm description

4.1.1. Data pre-processing

A considerable amount of missing and erroneous data were recorded and hence data pre-processing was performed prior to the application of the hybrid model. One can either remove the examples with erroneous data or remove the attributes with corrupted data. For the former, data processing was carried out using the Filter Examples operator of the RapidMiner Studio [5] software.[1]

1 The data were obtained from data.govt.uk [18]. The variable names have been changed, keeping the semantics same as before.

1 *Rapidminer* is an open source statistical and data mining tool.

This operator filters out the data entries according to the conditions specified by the user. Removing the attributes with corrupted data is achieved with the help of the Select Attributes operator in RapidMiner. The Filter Examples operator reduces the number of data entries in a dataset, but it has no effect on the number of attributes. On the other hand, the Select Attributes operator chooses the attributes with no missing or corrupted values and has no effect on the number of examples in the example set.

4.1.2. Data Re-sampling

The dataset used consists of 2044 data points for Class 1 accidents, 21098 data points for Class 2, and 93321 data points for Class 3, as shown in the form of a distribution curve in Figure 4. This accounts for the severe imbalance in the data, causing the prediction results to be skewed significantly in favor of the majority class. This causes poor classification rates on minor classes and extreme biasing towards the majority class. In addition, it is also possible that the classifier predicts everything as a major class and ignores the minor class. To tackle this issue, one must use re-sampling techniques to balance the data. We used the Synthetic Minority Oversampling (SMOTE) [9] up-sampling technique, which works by creating synthetic observations based upon the existing minority observations.

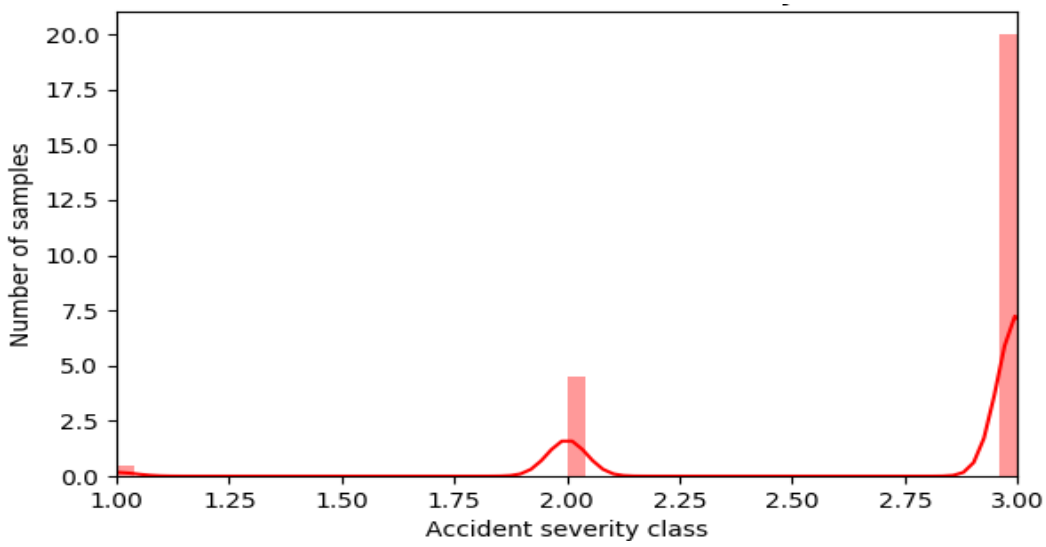

Figure. Accident severity class distribution

4.1.3. Feature/Attribute selection

In machine learning and statistics, feature selection or attribute selection is the process of selecting a subset of features that are used to build the model. It is performed to get rid of any unnecessary, irrelevant, or redundant features from the dataset, consequently resulting in improving the accuracy of the model. This also leads to better interpretability of the underlying relationship between input variables and the target class.

In this study, feature relevance analysis was carried out using the RapidMiner studio. The Weight by Tree Importance operator was used to find the relevant features. The weights of the attributes are calculated by analyzing the split points of a Random Forest model. Each node of each tree is visited, and the benefit created by the respective split is retrieved, which is further summed per attribute. The importance ranking is done by calculating the mean benefit over all the trees. This approach was implemented following the idea from the seminal work by Menze et al (2009) [12]. The higher the weight of the attributes is, the greater is their relevance. The Information Gain method was used to find the weights by the tree importance operator. Information Gain (IG) measures how much information a feature gives us about the class which is the entropy of the

distribution before the split minus the entropy of the distribution after it. Mathematically, the information gain is given by the equation below:

$$IG = E(p) - w * E(c) \quad (1)$$

where *IG* is information gain, *E* is entropy, *p* is parent node, c stands for children and w corresponds to the average of the weights.

4.1.4. Gaussian Mixture Model

Accident severity data can be formulated as a weighted sum of K-component Gaussian distributions:

$$p(x/\lambda) = \sum_{i=1}^{k} w_i N_i(x) \quad (2)$$

where x is a d-dimensional vector, $N_i(x)$ are the component multivariate Gaussian densities and wi is the mixing proportion or the mixture weights with $\sum_{i=1}^{k} w_i = 1$
Each component multivariate Gaussian density function is given by

$$N_i(x) = \frac{1}{\left((2\Pi)^{\frac{d}{2}}|\Sigma_i|^{-1/2}\right)} e^{\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)} \quad (3)$$

with μ_i , Σ_i as the mean vector and the Co-variance matrix respectively. The above-mentioned parameters, namely w_i , μ_i and Σ_i , are represented by

$$\lambda = (w_i, \mu_i, \Sigma_i) \; \forall i = 1,2,3,\ldots,K \quad (4)$$

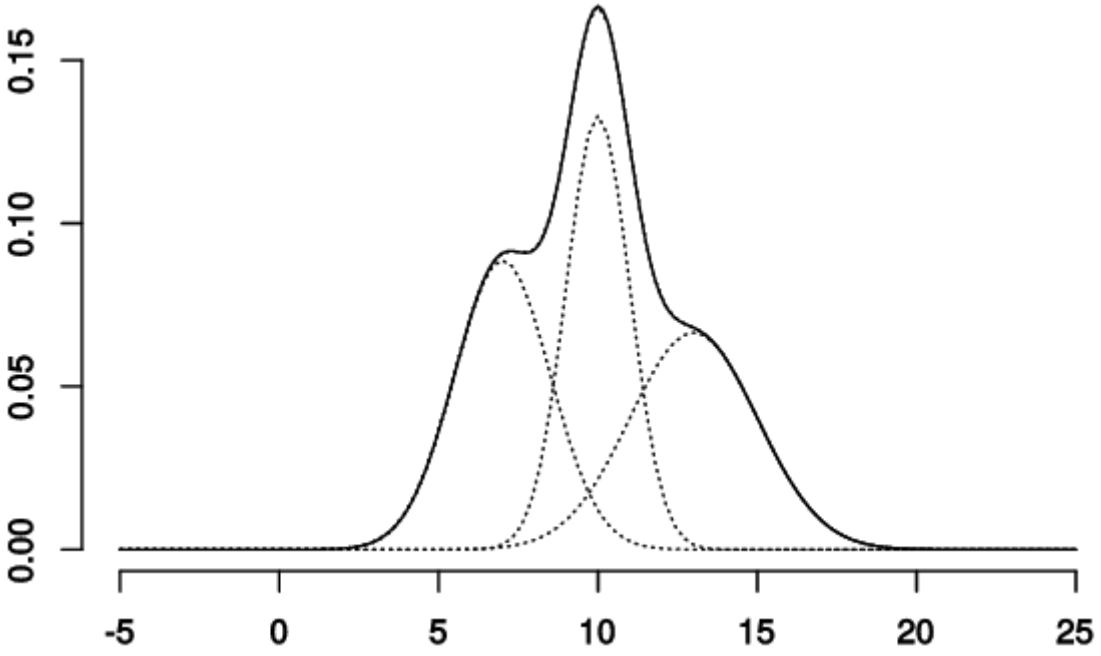

Figure 5: Gaussian mixture model with K=3

Given the M training vectors $x = (x_1, x_2, x_3, \ldots, x_M)$, the GMMs are trained with parameter evaluation using Maximum Likelihood (ML) estimation. Assuming all the training vectors are independent, the likelihood function and the log likelihood function turn out to be

$$p(x/\lambda) = \prod_{j=1}^{M} p\left(x_j/\lambda\right) \quad (5)$$

with the log likelihood using Equation 2 as

$$L(x/\lambda) = \sum_{j=1}^{M} log\left(\sum_{i=1}^{K} w_i N_i(x)\right) \quad (6)$$

The maximization of the likelihood function in Equation 5 is achieved by using the Expectation-Maximization (EM) algorithm. In the expectation (E) step, a function for the expectation of the log-likelihood is constructed while in the maximization (M) step, model parameters like the mean, variance and the mixing probability are estimated by the maximizing function found in the E step. After simplification, the formulas obtained performed at each E-M step are:

E step: Posterior probability estimation

$$p(i/x_i\,,\lambda) = \frac{w_i N_i(x_j)}{\sum_{l=1}^{K} w_l N_l(x_j)} \qquad (6)$$

M Step: Updating the parameters

$$w_i = \frac{1}{M}\sum_{j=1}^{M} p\left(i/x_j\,,\lambda\right) \qquad (8)$$

$$\mu_i = \frac{\sum_{j=1}^{M} p(i/x_j,\lambda)x_j}{\sum_{j=1}^{M} p(i/x_j,\lambda)} \qquad (9)$$

$$\Sigma_i = \frac{\sum_{j=1}^{M} p(i/x_j,\lambda){x_j}^2}{\sum_{j=1}^{M} p(i/x_j,\lambda)} - {\mu_i}^2 \qquad (10)$$

4.1.5. GMM and Traffic prediction

The observations (x) including features like weather conditions, lighting conditions, age of the driver, distance from junction etc., are assumed to be a mixture of three Gaussians (λ) which correspond to the three accident classes. Hence, our objective is to find a model that maximizes the posterior probability:

$$max_{1\le k\le K}\ \ p(\lambda_k/x), \qquad (11)$$

which by Bayes's rule is:

$$max_{1\le k\le K}\ \frac{p(\lambda_k/x)p(\lambda_k)}{p(x)}\,. \qquad (12)$$

Assuming all the Gaussians to be equally likely and taking the log, we have our likelihood function as:

$$max_{1\le k\le K}\ \sum_{j=1}^{M} logp\left(x_j/\lambda_k\right) \qquad (13)$$

which is further reduced to (6) and solved using expectation maximization.

4.1.6. Support vector classification

A support vector classifier or SVC is a discriminative model that makes decisions by constructing an optimal hyperplane or a line among linearly or non-linearly separable classes [14].

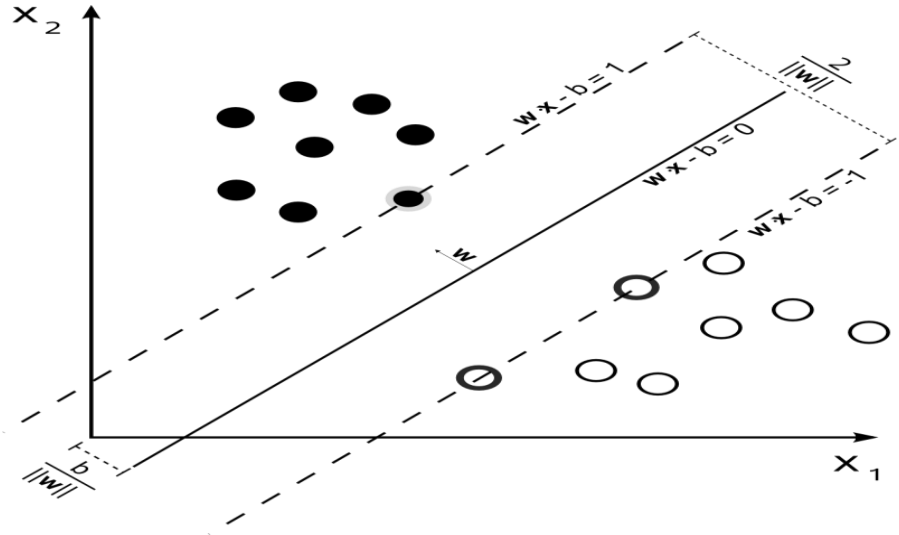

Figure 6: Concept of optimal hyper-plane

For linear support vector classifiers on the data (x_i, y_i) with i=1,2...n, the classification function is represented as:

$$f(x) = w^T(x) + b \qquad (14)$$

The margin according to Figure 6 is given by

$$\frac{|w^T(x)+b|}{\|w\|}\Big|_{w^T(x)+b=1} + \frac{|w^T(x)+b|}{\|w\|}\Big|_{w^T(x)+b=-1} = \frac{2}{w} \quad (15)$$

Since $w^T(x) + b = \pm 1$ for the support vectors.

Maximizing the margin (the minimum distance of the hyperplane from these points), the problem can be formulated as follows:

$$min \frac{1}{2}\|\mathrm{w}\|^2 \;\; s.t. \;\; y_i(x_i w + b) \geqslant 1 \qquad (16)$$

The solution for the optimal w turns out to be a linear combination of support vectors i.e. which satisfy $y_i(x_i w + b) = 1$.

In the case of a non-linearly separable dataset, no hyperplane exists that satisfies the above-mentioned constraints. In that case, a new model is introduced [13] :

$$min \frac{1}{2}\|\mathrm{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$s.t. \quad y_i(x_i w + b) \geqslant 1 - \xi_i \qquad i = 1,2,\ldots,n \qquad (17)$$

$$\xi_i \geqslant 0, \; i = 1,2,\ldots,n$$

where ξi is a non-negative factor called the slack variable responsible for allowing the functional value of certain samples to be negative. The factor 'C' is used to penalize the outliers and expresses the degree to which they are not acceptable.

The solution for the optimal w is a linear combination of all points $(\sum_{i=1} \alpha_i \, y_i x_i)$ in the feature space that have ξi > 0 and lie on the margin ($\alpha_i \neq 0$) and hence the classification function becomes:

$$f(x) = sign[(\sum_{i=1}^{n} \alpha_i \, y_i x_i)^T x + b] \qquad (18)$$

$$= sign[\sum_{i=1}^{n} \alpha_i \, y_i \langle x_i, x \rangle + b]$$

The non-linear classifier can be extended using the kernel function (K) satisfying Mercer's condition to map the input features to a higher dimensional space where it is linearly separable [15],

as represented in Figure 7. Then all the inner products are replaced with the kernel function and hence the classification function becomes,

$$f(x) = sign[\sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b] \qquad (19)$$

The most commonly used kernel functions are:

1. Polynomial kernel of degree d

$$K(x, y) = (\langle x, y \rangle + 1)^d \qquad (20)$$

2. Radial basis function (RBF)

$$K(x, y) = e^{\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)} \qquad (21)$$

3. Hyperbolic tangent (Sigmoid) kernel

$$K(x, y) = tanh(\alpha . \langle x, y \rangle + c) \quad (22)$$

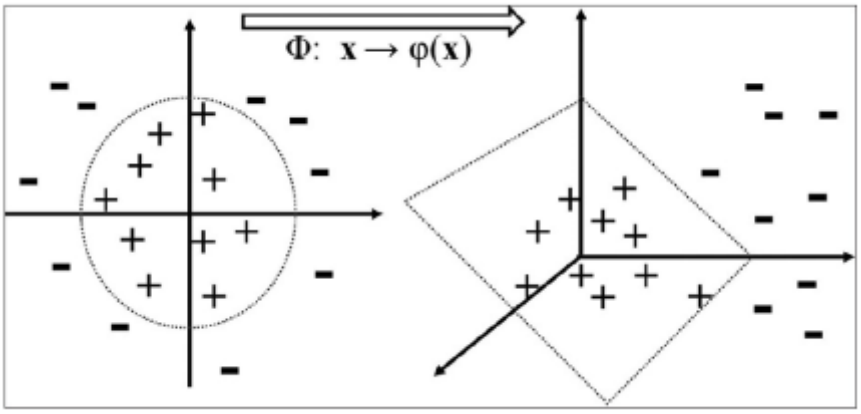

Figure 7. Kernel trick

4.1.7. Multiclass SVC

1. One-against-all method:
This method [16] considers N classifiers where N is the number of classes and trains the i^{th} classifier with all other examples considering the instances of the i^{th} class as positive and all other instances as negative labels.

2. One-against-one method:
This method [17] constructs N (N − 1)/2 classifiers and trains the i^{th} classifier with every j^{th} classifier considering the instances of the i^{th} classifier as positive and those of the j^{th} classifier as negative.

4.2. The Need for a Hybrid Model

Since research has shown that no single method is best for every situation in traffic prediction, building hybrid models (HMs) is an approach that combines different methods that could produce better results than any of those methods applied individually.

In this paper, the need for HMs is justified using a statistical modeling method i.e., a Gaussian Mixture Model (GMM) and a machine learning modeling scheme i.e., the Support Vector Classifier (SVC). As the baseline generative model (GMM) could only classify with maximum likelihood, the SVM presents an attractive way of enhancing it. This is because of the SVM's innate discriminative power, even in the case of non-linearly separable classes using kernels. However, SVCs also perform poorly on unbalanced data. Hence the GMM serves as a parametric basis for the support vector classifier. Therefore, in this work, the use of a hybrid model (HM) was needed in

order to overcome the disadvantages of one model by the advantages of the other and hence to improve the accuracy.

5. DISCUSSION OF RESULTS

5.1. Data pre-processing results

Erroneous and missing data entries were removed using the RapidMiner Studio. The data was reduced from 178918 examples and 69 attributes to 116463 examples and 69 attributes, removing all the missing values. By removing the attributes with missing values, the number of features was reduced from 69 to 62. The variables and the number of missing values are listed in Table IV below.

Variable name	**Number of missing values**
Index of the crash	53701
LSOA of crash location	17736
Longitude	59
Latitude	59
Location Easting OSGR	35
Location Northing OSGR	35
Time	3

Table IV: Attributes with missing and erroneous values

5.2. Data Re-sampling results

The imbalanced dataset is balanced using SMOTE from the 'imblearn' module of python. This is an upsampling technique which balances the data by increasing the number of data points for the minority class. After applying SMOTE on our dataset, we received a total of 279,963 samples with 93,170 samples from Class 1, 93,756 samples from Class 2 and 93,037 from Class 3, as listed
in Table V.

Accident severity class	**Training samples before SMOTE**	**Training samples after SMOTE**
Class 1	2,044	93,170
Class 2	21,098	93,756
Class 3	93,321	93,037

Table V: Data re-sampling results

5.3. Feature selection results

The variable importance ranking based on the three accident severity levels was conducted using the RapidMiner Studio. Weight by the tree importance ranking operator was used after applying the Random forest model on the processed data. In addition to this, the information gain method was used to find weights using tree importance. In this, the variable with the largest score is normalized to 1 and the scores of all the others are calculated with respect to the best performing variable.

The results obtained are shown below in Table VI and Figure 8. It can be seen that among all the variables, Intensity of the fatality is the most related to accident severity with a score of 1. The location attributes like Location Northing OSGR, Latitude, Location Easting OSGR, Longitude

follow in the list. The Pedestrian Road Maintenance Worker variable was of the least importance with a score of 0.0. Surprisingly, weather conditions had a relevance of 4.2%, which is quite insignificant. Factors like maximum velocity possible, day of the week, vehicles involved in the accident, date, etc. contributed significantly with a score above 20%. Features like details of intersection, location of intersection, kind of road, lighting conditions, age of the casualty, etc., also turn out to be quite important to the hybrid model. All features varying from environmental, physical, geometrical, geographical and historical were included in the top features ranked using this technique. The results obtained are in accordance with one's personal experience and knowledge about the risk factors related to accidents.

Variable	**Score**	**Variable**	**Score**	**Variable**	**Score**
Intensity of the fatality	1	Age of victim	0.092	Did a Police Officer attend scene of accident?	0.042
Location northing OSGR latitude	0.699	Lighting	0.091	Type of area	0.036
Latitude	0.693	Location of intersection	0.086	Hit object off carriage way	0.029
Longitude	0.661	Location of victim's house	0.078	Type of fatality	0.027
Location easting OSGR	0.654	Kind of road	0.075	Vehicle Id code x	0.023
Date	0.439	Age band of victim	0.073	Gender of the driver	0.021
1st Road Number	0.385	Kind of road	0.069	Rider home area type	0.018
District of local authority	0.317	Skidding and overturning	0.067	Vehicle Id code y	0.018
Vehicles involved	0.288	Carriageway hazards	0.067	Casualty Id code	0.017
Day of week	0.262	Road surface conditions	0.063	Special characteristics of accident location	0.016
Area of police responsible	0.222	2nd road class	0.063	Gender of victim	0.014
Maximum velocity possible	0.21	Vehicle IMD Decile	0.062	Position of the pedestrian	0.012
Number of victims involved	0.21	1st point of impact	0.06	Hit object in Carriageway	0.011
Details of intersection	0.175	Causalty IMD Decile	0.055	Propulsion code	0.011
1st Road Class	0.129	Age band of the driver	0.052	Towing and articulation	0.01
Age of the vehicle?	0.12	Journey purpose of driver	0.052	Position of victim in Bus/Coach	0.01
Age of the driver	0.117	Type of victim	0.051	Position of vehicle	0.01
Traffic control at intersection	0.113	Pedestrian crossing-physical facilities	0.05	Motion of the pedestrian	0.009
Rider IMD Decile	0.11	Vehicle Manoeuvre	0.049	Was vehicle left hand drive?	0.005
Vehicle leaving carriageway	0.109	Pedestrian crossing-human control	0.046	Pedestrian road maintenance worker	0
2nd Road Number	0.108	Position of victim in car	0.045		
Motor power (CC)	0.106	Weather conditions	0.042		

Table VI. Variable relevance scores

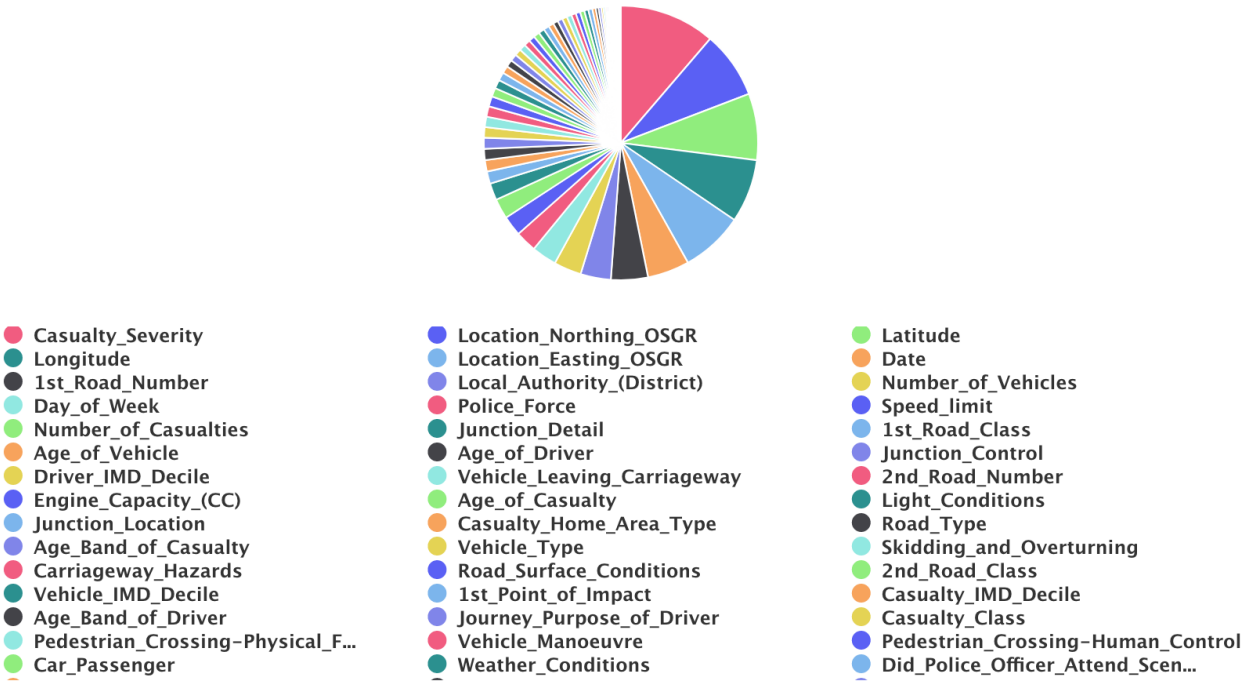

Figure 8: Variable importance score distribution

Accident class	**Mean matrix**
Class 1	[2.5, 162471.58, 51.35, -0.53, 502646.61, 381.98, 460.20, 3.60, 4.03, 44.59, 48.21, 3.11, 1.10, 3.13, 5.64, 1.32, 0.42, 4.34, 0.60, 305.99, 1708.36, 40.26, 2.19, 1.15, 0.62]
Class 2	[2.43, 119903.62, 50.98, -1.17, 458311.88, 400.35, 497.09, 3.12, 44.00, 55.88, 3.40, 0.99, 2.06, 7.31, 1.36, 0.22, 5.66, 0.48, 0.00, 1733.54, 43.35, 3.42, 1.53, 1.19]
Class 3	[2.59, 159126.65, 51.30, 0.74, 591387.43, 234.60, 537.71, 2.85, 7.29, 46.00, 90.77, 2.94, 1.17, 3.98, 7.41, 1.29, 0.34, 4.70, 0.48, 8.20, 1643.93, 35.72, 2.79, 0.88, 1.13]

Table VII. Mean vectors using Gaussian mixture model

5.4. Hybrid Gaussian mixture model and support vector classifier results

After data pre-processing and re-sampling, 120,00data samples with 39.996 samples from Class 1, 39,998 samples from Class 2, and 40,006 samples from Class 3 were used as input to the Gaussian mixture model. The top 25 features according to the variable importance ranking results were chosen as features of the input data entries. The data were fitted with a mixture of three Gaussians which correspond to the three accident severity classes. Moreover, principal component analysis (PCA) was applied to visualize the results in 2 dimensions.

The results with clustering based on the predictions by the Gaussian mixture model are shown in Figure 9. The mean matrices obtained for all three classes are also listed in Table VII.

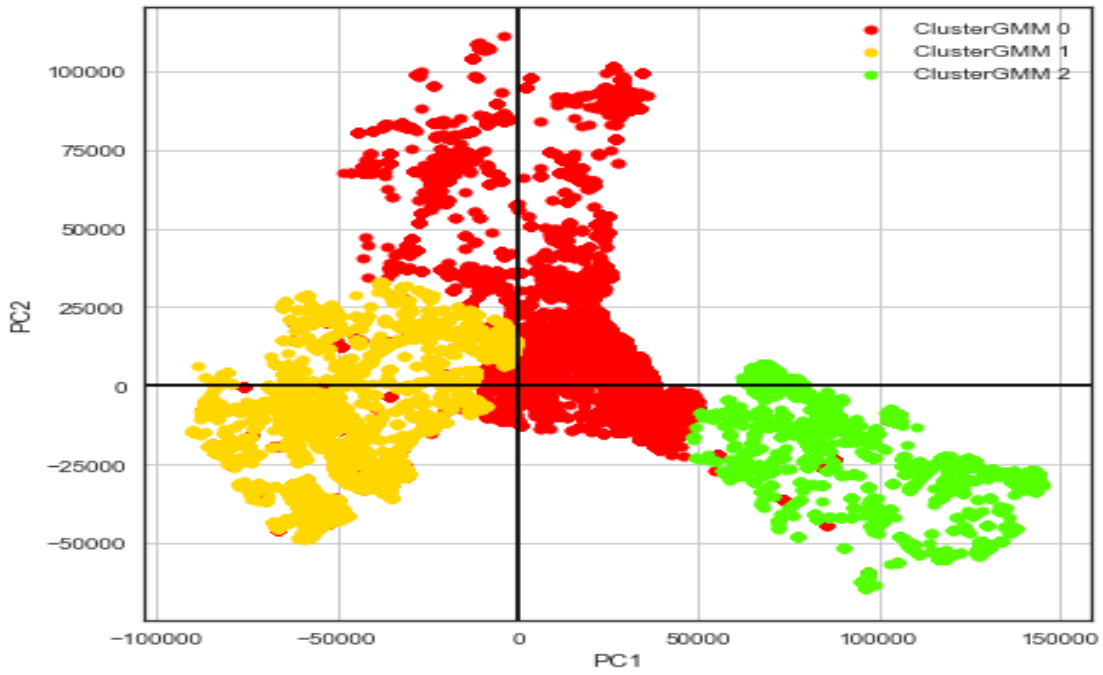

Figure 9. Gaussian mixture modelling results

These results suggest that for a particular class of accident severity, the mean vector is the average value of the observed features. To interpret these results, let us consider an example: if a person is driving at a longitude of -0.527 and has a speed of nearly 48.2 km/hr., then he/she is very likely to have an accident of Class 1 severity, i.e. a no-injury accident. On the other hand, a person traveling at a longitude of 0.743 with a speed of 90.7 km/hr. has very high chance of having an accident of Class 3 severity, i.e. an incapacitating injury. Similarly, with respect to the day of the week, the mean value observed rounded to the nearest integer for Classes 1, 2 and 3 are 4, 3 and 7 respectively. This implies that accidents of the highest severity or fatal accidents are likely to occur on Saturdays. One possible reason for this could be that on Saturdays, there are more cars and more drivers are impaired by alcohol due to weekend celebrations and parties than on any other day. These results are fairly consistent with one's personal experience and logical reasoning.

It is also observed that some values in the mean vectors of Classes 2 and 3 are very close to each other. The reason for this is the uneven distribution of the values inside every feature. For example, the feature, 'Lighting' has 126,049 datat points corresponding to daylight while only 1,123 and 10,314 for darkness light unlit and darkness -no lighting respectively. Thus, daylight itself accounts for 70.45% of the example set, which is a very high number. This sub-skewing of data leads to biasing in favor of the majority class.

The overall accuracy of the Gaussian Mixture Model was 64.68%. These 3 mean vectors were used as input to the support vector classifier. For such a large dataset with 120,000 examples, 3 data points for training would be insufficient and would lead to overfitting. As a result, we used some extra data points alongside the mean vectors for the purpose of training the support vector classifier and further decreasing the testing data. Had there been more classes of accident severity in the dataset, one could have directly used the mean vectors as input for the SVC and hence improve the model performance, like the technique applied in text-independent speaker identification using both SVM and GMM. [8].

The SVC with radial basis function produced a total accuracy of 84.35%. Precision recall and the F1-score were calculated to quantify the performance of our classifier. In order to compute these parameters, 4 performance metrics given below are evaluated from the confusion metric:

- True Positives (TP) - These are the examples with ’yes’ as their actual class as well as the class predicted by the model.
- True Negatives (TN) - These are the examples with ’no’ as their actual class as well as the class predicted by the model.
- False Positives (FP) - These are the examples with ’no’ as their actual class but are predicted as ’yes’ by the model.
- False Negatives (FN) - These are the examples with ’yes’ as their actual class but are predicted as ’no’ by the model.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Table VIII. Confusion metric

Subsequently, the performance estimation parameters are defined as

- Accuracy (A): This is defined as the ratio of the number of correctly predicted examples over the total number of examples. Hence, we have :

$$\frac{TP+TN}{TP+FP+FN+TN}$$

- Precision (P): This is defined as the ratio of correctly predicted positive observations to the total predicted positive observations. We then have:

$$\frac{TP}{TP+FP}$$

– Macro precision: Precision found by calculating metrics for each label, and then finding their un-weighted mean.

– Micro precision: Precision found by calculating metrics globally by counting the total number of true positives, false negatives and false positives.

- Recall (Sensitivity) (R): Recall is the ratio of correctly predicted positive observations to the all positive observations in actual class, which means:

$$\frac{TP}{TP+FN}$$

– Macro Recall: Recall found by calculating metrics for each label, and then finding their un-weighted mean.

– Micro Recall: Recall found by calculating metrics globally by counting the total number of true positives, false negatives and false positives.

- F1 score: F1 Score is the weighted average of Precision and Recall and is used to combine precision and recall in a single metric as in the following: $2 * \frac{P*R}{P+R}$

The performance scores including precision, recall and f1-score are listed below in Table IX. The radial basis function or the RBF achieved an accuracy of 88.52%, outperforming the linear kernel, which was 59.89% accurate.

Class	**Precision**	**Recall**	**f1-score**
Class 1	1.00	0.9342	0.9659
Class 2	1.00	0.7214	0.8381
Class 3	0.7437	1.00	0.8530
Macro average	91.4595%	88.5220%	88.5748%

Micro average	88.5166%	88.5166%	87.2895%
Weighted average	91.4588%	88.5166%	88.5711%
AUC	0.99	0.97	0.97
Accuracy	88.5167%		

Table IX. RBF kernel performance metrics

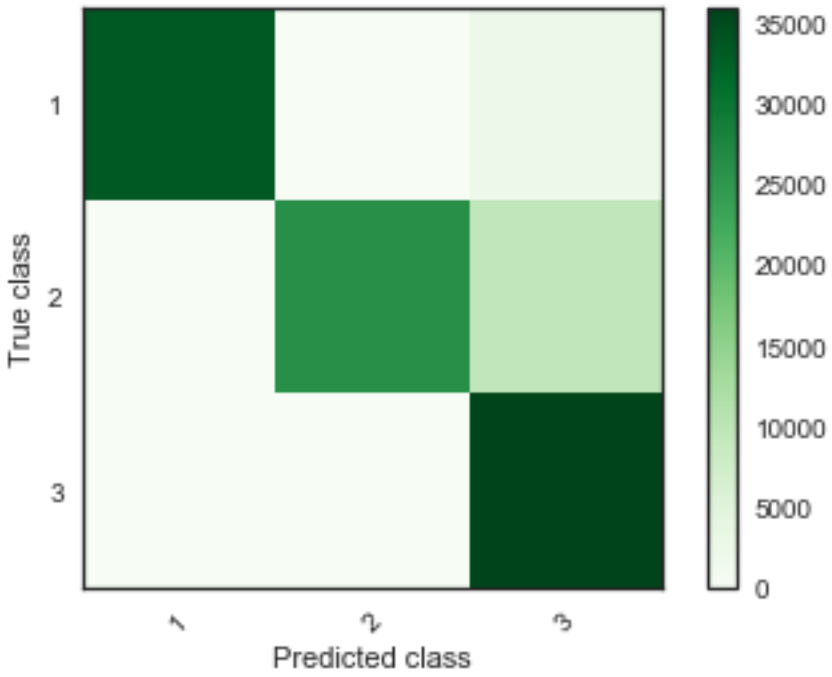

Figure 10. Confusion matrix for accident dataset

Figure 10 shows the confusion matrix obtained with the hybrid model. As can be observed, there is a clear separation between accidents without any injury (Class 1) and accidents with injury (Classes 2 and 3). Most of the confusion occurs between non-incapacitating injury accident and incapacitating injury accidents. Furthermore, the ROC (Receiver Operating Characteristics) curve and AUC/AUROC (Area Under the Receiver Operating Characteristics) were determined using the above parameters. ROC is a probability curve with TPR (y) plotted against the FPR (x) which is FP/TN+FP. The area under the ROC curve quantifies the model's ability to identify the classes correctly and distinguish between them [9]. The AUC-ROC curve for this model is shown in Figure 11. The AUC values for Classes 1, 2 and 3 are 0.99, 0.97 and 0.97 respectively. These values are very close to 1 and reflect the good discriminative power of the classifier.

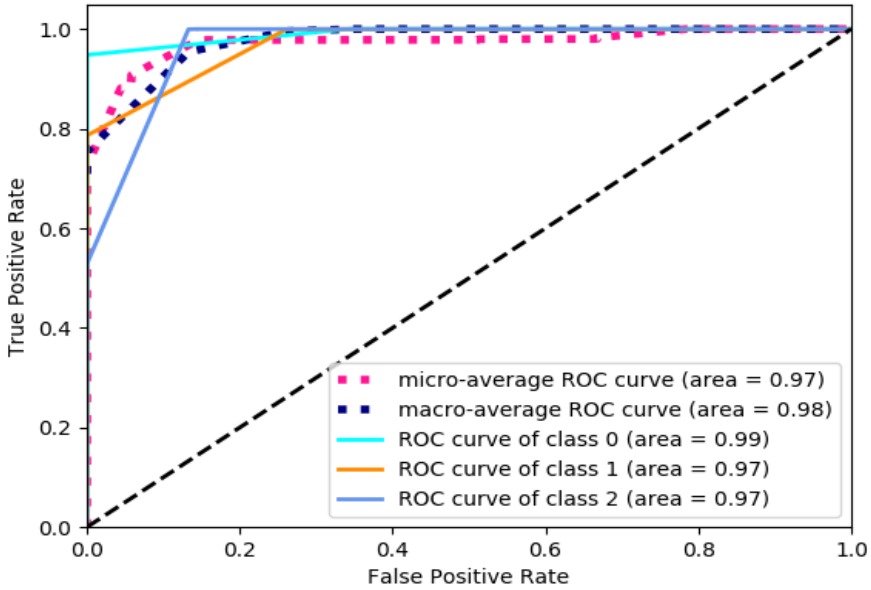

Figure 11. ROC curve for the accident dataset

Similarly to AUC-ROC, an area under precision/recall curve (AUC-PR) can also be calculated to show the tradeoff between precision and recall as a function of varying a decision threshold. The higher the area under the curve is, the higher are the values of precision and recall, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate [54]. For the hybrid model, the AUC-PR curve micro averaged over all classes is
shown in Figure 12. The AUC-PR curves for each class represented over the iso-F1 curves are plotted in Figure 13 where an iso-F1 curve is a curve containing all the points in the precision-recall space whose F1 scores are the same.

Figure 12. AUC-PR curve micro averaged over all classes for the accident dataset

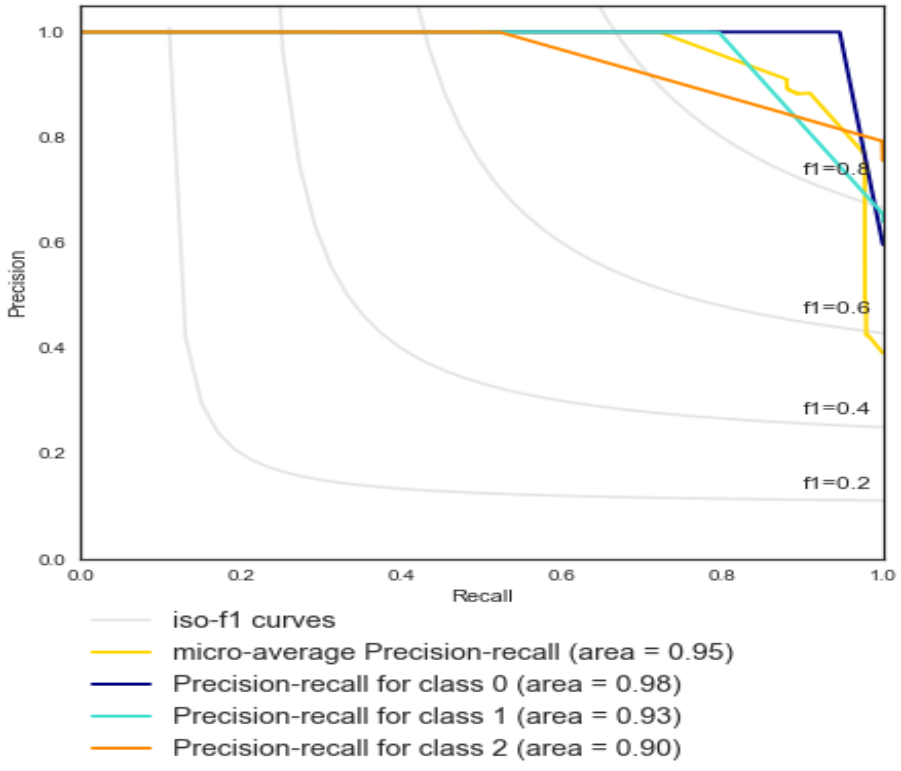

Figure 13. AUC-PR curves for each class on various ISO-F1 curves for the accident dataset

6. CONCLUSIONS

Road traffic accidents have become a major cause of injury and death. With increasing urbanization and growing populations, the volume of vehicles has increased exponentially. As a result, traffic accident forecasting, and the identification of accident-prone areas can help reduce the risks of traffic accidents and improve overall life expectancy.

The data about the circumstances of personal injury in road accidents, the types of vehicles involved, and the consequential casualties were obtained from data.govt.uk [18]. The output or the accident severity class was divided into three major categories namely: no injury in the accident, a non-incapacitating injury in the accident and an incapacitating injury in the accident. In this paper, a hybrid classifier was proposed which combines the descriptive strength of the baseline Gaussian Mixture Model (GMM) with the high-performance classification capabilities of the Support Vector Classifier (SVC). A new approach was introduced using the mean vectors obtained from the GMM model as input to the SVC. The model was supported with data pre-processing and re-sampling to convert the data points into suitable form and avoid any kind of biasing in the results. Feature importance ranking was also performed to choose relevant attributes with respect to accident severity. This hybrid model successfully took advantage of both models and obtained a better accuracy than the baseline GMM model. The radial basis kernel outperformed the linear kernel by achieving an accuracy of 85.53%. Data analytics performed including the area under the receiver operating characteristics curve (AUC-ROC) and the area under the precision/recall curve (AUC-PR) indicates the successful application of this model in traffic accident forecasting.

Although a significant improvement in accuracy has been observed, this study has several limitations. The first concerns the dataset used. This research is based on a road traffic accident dataset from the year 2017 which contains very few data samples for the no injury and non-incapacitating injury types of accident. The data was unbalanced not just with respect to the output class but also with respect to the sub features of various attributes. Moreover, aggregating the accident severity into just three categories limits the scope of the study and the results obtained. The greater the number of severity classes, the less is the amount of extra training data required to feed into the SVC to avoid overfitting. Thus, datasets with enough records corresponding to each class are desirable and should be used for in further studies.

The second limitation concerns the dependence of the SVC model on parameters and attribute selection. In this study, the performance of SVC relies heavily on the feature selection results and the mean vectors obtained from the GMM. In order to improve the accuracy of the support vector classifier, other approaches like particle swarm optimization (PSO), ant colony optimization, genetic algorithms, etc. could be used for effective parameter selection. In addition to this, more kernels like the polynomial kernel and the sigmoid kernel could be tested in order to improve future model performances.

References

[1] M. Jeong, B. C. Ko, and J.-Y. Nam, Early detection of sudden pedestrian crossing for safe driving during summer nights, IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 6, pp. 13681380, 2017.

[2] C. Dong, C. Shao, J. Li and .Z. Xiong1, ”An Improved Deep Learning Model for Traffic Crash Prediction,” Journal of Advanced Transportation,vol. 2018, Article ID 3869106, https://doi.org/10.1155/2018/3869106

[3] American Association of state Highway and Transportation Ofcials ”Highway Safety Manual.”, Washington, D.C. (2010).

[4] C, Chen, G. Zhanga, R. Tarefder, J. Ma, H. Wei and H. Guand, ”A multinomial logit model-Bayesian network hybrid approach for driver injury severity analyses in rear-end crashes,” Accid. Anal. Prev., vol. 80, 2015, pp.76-88.

[5] I. Mierswa, and R. Klinkenberg,science,machine learning, RapidMiner Studio (9.1) [Data predictiveanalytics].Retrieved from https://rapidminer.com/

[6] S. Jin, X. Qu and D. Wang, ”Assessment of Expressway Traffic Safety Using Gaussian Mixture Model based on Time to Collision,” International Journal of Computational Intelligence Systems., vol. 4, No. 6, pp. 1122-1130, Dec., 2011.

[7] S. Jin, D. Wang, C. Xu and D. Ma, ”Short-term traffic safety forecasting using Gaussian mixture model and Kalman filter,” Journal of Zhejiang University SCIENCE A., vol.14, Issue 4, pp. 231243, Apr., 2013.

[8] H.Bourouba, C.A.Korba, R. Djemili, ”Novel Approach in Speaker Identification using SVM and GMM,” Journal of control engineering and applied Informatics (CEAI)., vol.15, no.3 pp. 87-95, 013

[9] H.Bourouba, C.A.Korba and R. Djemili, ”SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research 16.pp. 321357, 2002

35[10] Akoz and M.E. Karsligil, ”Severity detection of traffic accidents at intersections based on vehicle motion analysis and multiphase linear regression,”Journal of Intelligent Transportation Systems., 2010.

[11] S. Sun, C. Zhang, and G.Yu, ”A Bayesian Network Approach to Traffic Flow Forecasting,” Journal of Intelligent transportation systems., vol. 7, no. 1, Mar. 2006.

[12] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich and F. A. Hamprecht, ”A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data,” BMC Bioinformatics 2009.,10:213, doi:10.1186/1471-2105-10-213

[13] C. Cortes and V. Vapnik,” Machine Learning (1995)., 20: 273., doi:10.1023/A:1022627411411

[14] J. A. Suykens and J. Vandewalle, Least squares support vector machine classifiers, Neural processing letters., vol. 9, no. 3, pp. 293300, 1999.

[15] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[16] C. M. Bishop, Pattern recognition and machine learning., Springer (2006).

[17] S. Knerr, L. Personnaz, and G. Dreyfus, ”Single-layer learning revisited: a step-wise procedure for building and training a neural network.”, Neurocomputing: Algorithms, Architectures and Applications., Springer-Verlag, Vol.F68, p. 41-50, 1990.

[18] Department for Transport, ’Road Safety Data’, 2014. [Online]. Available: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.

[19] J. Bilmes, A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, Tech. Rep. ICSI-TR-97-021, University of California Berkeley, April 1998.

36[20] T. K. Moon, The expectation-maximization algorithm, IEEE Signal Processing Magazine, 13(6):7460, November 1996.
[21] Jain, A.K.; Duin, R. P W; Jianchang Mao, ”Statistical pattern recognition: a review,” Pattern Analysis and Machine Intelligence, IEEE Transactions, vol.22, no.1, pp. 4-37, Jan. 2000.
[22] E. Hauer, J.C.N. Ng and J. Lovell, ”Estimation of safety at signalized intersections,” Transp. Res. Rec., vol.1185, pp. 48-61, 1988.
[23] R.L. Cheu, J. Xu, A.G.H. Kek, W.P. Lim and W.L. Chen, ”Forecasting of shared-use vehicle trips using neural networks and support vector machines,” Transp. Res. Rec., vol.1968, pp. 40-46, 2006.
[24] M.-L. Huang, ”Intersection traffic flow forecasting based on -GSVR with a new hybrid evolutionary algorithm,” Neurocomputing., vol.147, pp. 343-349, 2015.
[25] D. Wei and H. Liu, ”An adaptive-margin support vector regression for short-term traffic flow forecast,” Intelligent Transportation Systems.,vol.17, pp. 317-327, 2013.
[26] R. Yu, G. Wang, J. Zheng and H. Wang, ”Urban Road Traffic Condition Pattern Recognition Based on Support Vector Machine,” Journal of Intelligent Transportation Systems., vol.13, pp. 130-136, 2013.
[27] X. Li, D. Lord, Y. Zhang and Y. Xie, ”Predicting motor vehicle crashes using Support Vector Machine models,” Accid. Anal. Prev., vol.40, pp. 1611-1618, 2008.
[28] G. Ren and Z. Zhou, ”Traffic safety forecasting method by particle swarm optimization and support vector machine,” Expert Systems with Applications., vol.38, pp. 10420-10424, 2011
[29] A. Surez Snchez, P. Riesgo Fernndez, F. Snchez Lasheras, F.J. de Cos Juez and P.J. Garca Nieto, ”Prediction of work-related accidents according to working conditions using support vector machines,” Applied Mathematics and Computation., vol.218, pp. 3539-3552, 2011
[30] R. Yu and M. Abdel-Aty, ”Utilizing support vector machine in real-time crash risk evaluation,” Accid. Anal. Prev., vol.51, pp. 252-259, 2013.
[31] L. Guo, P.-S. Ge, M.-H. Zhang, L.-H. Li and Y.-B. Zhao, ”Pedestrian detection for intelligent transportation systems combining AdaBoost algorithm and support vector machine,” Expert Systems with Applications., vol.39, pp. 4274-4286, 2012.
[32] R. Yu and M. Abdel-Aty, ”Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data,” Saf. Sci., vol.63, pp. 50-56, 2014.
[33] C Chen, G Zhang, Z Qian, RA Tarefder and Z.Tian, ”Investigating driver injury severity patterns in rollover crashes using support vector machine models” Accid. Anal. Prev., vol. 90, pp. 128-139, May 2016.
[34] J.D. Oa, G. Lpez and J. Abelln, ”Extracting decision rules from police accident reports through decision trees,” Accid. Anal. Prev., vol.50, pp. 1151-1160, 2013.
[35] J. Abelln, G. Lpez and J.D. Oa, ”Analysis of traffic accident severity using decision rules via decision trees,” Expert Systems with Applications., vol.40,pp. 6047-6054, 2013.
[36] H.T. Abdelwahab and M.A. Abdel-Aty, ”Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections,” Transp. Res. Rec., vol.1746, pp. 6-13, 2001
[37] D. Delen, R. Sharda and M. Bessonov, ”Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks,” Accid. Anal. Prev., vol.38, pp. 434-444, 2006
38[38] Q. Zeng and H. Huang, ”A stable and optimized neural network model for crash injury severity prediction,” Accid. Anal. Prev., vol.73, pp. 351-358, 2014.
[39] A. Das, M. Abdel-Aty and A. Pande, ”Using conditional inference forests to identify the factors affecting crash severity on arterial corridors,” Journal of Safety Research., vol.40, pp. 317-327
[40] R. Harb, X. Yan, E. Radwan and X. Su, ”Exploring precrash maneuvers using classification trees and random forests,” Accid. Anal. Prev., vol.41, pp. 98-107, 2009
[41] T.K. Anderson, ”Kernel density estimation and k-means clustering to profile road accident hotspots,” Accid. Anal. Prev., vol.41, pp. 359-364, 2009.

[42] R. Mauro, M. Luca and G. Acqua, ”Using a K-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis,” Mod. Appl. Sci., vol.7, pp. 11-19, 2013.
[43] N. Dong, H. Huang and L. Zheng, ”Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects,” Accid. Anal. Prev., vol.82, pp. 192-198, 2015.
[44] Z. Li, P. Liu, W. Wang and C. Xu, ”Using support vector machine models for crash injury severity analysis,” Accid. Anal. Prev., vol.45, pp. 478-486, 2012.
[45] A. Iranitalab and A. Khattak, ”Comparison of four statistical and machine learning methods for crash severity prediction,” Accid. Anal. Prev., vol.108, pp. 27-36, 2017.
[46] D. Mahalel, ”A note on accident risk,” Transp. Res. Rec., vol.1068, pp. 85-89, 1986.
[47] J.Tang, J. Liang, C. Han, Z. Li and H. Huang, ”Crash injury severity analysis using a two-layer Stacking framework,” Accid. Anal. Prev., vol.122, pp. 226-238, January 2019.
[48] J. Tang, F. Liu, W. Zhang, R. Ke and Y. Zou, ”Lane-changes prediction based on adaptive fuzzy neural network,” Expert Systems With Applications., vol.91, pp. 452-463, January 2018.
[49] C. Chen, G. Zhang, R. Tarefder, J. Ma, H. Wei and H. Guan, ”A multinomial logit model-bayesian network hybrid approach for driver injury severity analyses in rear-end crashes,” Accid. Anal. Prev., vol.90, pp.128-139,May 2016.
[50] C. Chen, G. Zhang, Z. Tian, S.M. Bogus and Y. Yang, ”Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations,” Accid. Anal. Prev.,vol.85, pp. 186-198, 2015.
[51] J. Liu, X. Wang, A.J. Khattak, J. Hu, J. Cui and J. Ma, ”How big data serves for freight safety management at highway-rail grade crossings: a spatial approach fused with path analysis,” Neurocomputing., vol.181, pp.38-52, 2015.
[52] D. Lord and F. Mannering, ”The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives,” Transportation Research Part A., vol.44, pp. 291-305, 2010.
[53] Q. Wu, G. Zhang, Y. Ci, L. Wu, R. A. Tarefder and A. Alcntara, ”Exploratory multinomial logit model-based driver injury severity analyses forteenage and adult drivers in intersection-related crashes,” Traffic Injury Prevention., vol.17, pp. 413-422, 2015.
[54] J. Davis and M. Goadrich, The relationship between Precision-Recall and ROC curves, Proc. 23rd ICML, 2006, pp. 233240.
[55] Chang, L.Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf. Sci. 43 (8), 541557.
[56] Guangyuan Pan , Liping Fu, Lalita Thakali. Development of a global road safety performance function using deep neural networks. International Journal of Transportation Science and Technology journal homepage: www.elsevier.com/locate/ijtst Journal Article.
[57] Chang, L.Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Saf. Sci. 43 (8), 541557.
[58] S. Al-Sultan, M. M. Al-Doori, A. H. Al-Bayatti, and H. Zedan, A comprehensive survey on vehicular ad hoc network, Journal of network and computer applications, vol. 37, pp. 380392, 2014.
[59] Iqbal, Zafar and Khan, Majid Iqbal. Automatic incident detection in smart city using multiple traffic flow parameters via V2X communication. International journal of distributed sensor networks,volume,14 number(11) pp1550147718815845,2018, publisher=SAGE Publications Sage UK: London, England
[60] Wang J, Deng W and Guo Y. New Bayesian combination method for short-term traffic flow forecasting. Transport Res C: Emer 2014; 43(1): 7994.
[61] Sun S, Zhang C and Yu G. A Bayesian network approach to traffic flow forecasting. IEEE T Intell Transp 2006; 7(1): 124132
[62] Lawrence MJ, Edmundson RH and OConnor MJ. The accuracy of combining judgemental and statistical forecasts. Manag Sci 1986; 32(12): 15211532.
[63] Makridakis S. Why combining works? Int J Forecast 1989; 5(4): 601603.

[64] Jiaming Xie and Yi-King Choi: Hybrid traffic prediction scheme for intelligent transportation systems based on historical and real-time data. International Journal of Distributed Sensor Networks. 31 October 2017.
[65] M. Van Der Voort , M.Dougherty ,S. Watson , Combining Kohonen maps with Arima time series models to forecast traffic flow, Transp. Res. Part C Emerg.Technol. 4 (5) (1996) 307318 .
41[66] Z. Li , S. Jiang , L. Li , Y. Li , Building sparse models for traffic flow prediction: an empirical comparison between statistical heuristics and geometric heuristics for Bayesian network approaches, Transp. B Transp. Dyn. (2017) 117 .
[67] D. Huang, Z. Deng, L. Zhao, B. Mi, A short-term traffic flow forecasting method based on Markov chain and grey Verhulst model, in: Proceedings of the Data Driven Control and Learning Systems, 2017, pp. 606610.
[68] K.-L. Li , C.-J. Zhai , J.-M. Xu , Short-term traffic flow prediction using a methodology based on Arima and RBN-ANN, in: Proceedings of the Chinese Automation Congress, 2017, pp. 28042807 .
[69] M. Castro-Neto , Y.-S. Jeong , M.-K. Jeong , L.D. Han , Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions, Expert Syst. Appl. 36 (3) (2009) 61646173 .
[70] M. Shuai , K. Xie , W. Pu , G. Song , X. Ma , An online approach based
on locally weighted learning for short-term traffic flow prediction, in: Proceedings of the ACM SIGSPATIAL international conference on Advances in geographic information systems, 2008, pp. 4553.
[71] M. Lippi , M. Bertini , P. Frasconi , Collective traffic forecasting, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2010, pp. 259273
[72] Qinglu MA, Weining LIU,Dihua SUN, Saleem-Ullah Lar. Traffic Condition On-line Estimation Using Multi-source Data. Journal of Computational Information Systems 8: 6 (2012) 26272635 Available at http://www.Jofcis.com